Received: 27 June 2018 | Revised: 10 December 2018 | Accepted: 28 December 2018

DOI: 10.1002/ece3.4919

WILEY Ecology and Evolution Open Access

ORIGINAL RESEARCH

Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem

Hailu Shiferaw[1,2] | Woldeamlak Bewket[2] | Sandra Eckert[3]

[1]Water and Land Resource Centre, Addis Ababa University, Addis Ababa, Ethiopia

[2]Department of Geography and Environmental Studies, Addis Ababa University, Addis Ababa, Ethiopia

[3]Centre for Development and Environment, University of Bern, Bern, Switzerland

Correspondence
Hailu Shiferaw, Water and Land Resource Centre, Addis Ababa University, Addis Ababa, Ethiopia.
Emails: hailu2nd@gmail.com; hailushi31@yahoo.com

Funding information
Swiss National Science Foundation (SNSF); Swiss Agency for Development and Cooperation; Swiss Programme for Research on Global Issues for Development (r4d)

Abstract

In recent years, an increasing number of distribution maps of invasive alien plant species (IAPS) have been published using different machine learning algorithms (MLAs). However, for designing spatially explicit management strategies, distribution maps should include information on the local cover/abundance of the IAPS. This study compares the performances of five MLAs: gradient boosting machine in two different implementations, random forest, support vector machine and deep learning neural network, one ensemble model and a generalized linear model; thereby identifying the best-performing ones in mapping the fractional cover/abundance and distribution of IPAS, in this case called *Prosopis juliflora (SW. DC.)*. Field level *Prosopis* cover and spatial datasets of seventeen biophysical and anthropogenic variables were collected, processed, and used to train and validate the algorithms so as to generate fractional cover maps of *Prosopis* in the dryland ecosystem of the Afar Region, Ethiopia. Out of the seven tested algorithms, random forest performed the best with an accuracy of 92% and sensitivity and specificity >0.89. The next best-performing algorithms were the ensemble model and gradient boosting machine with an accuracy of 89% and 88%, respectively. The other tested algorithms achieved comparably low performances. The strong explanatory variables for *Prosopis* distributions in all models were NDVI, elevation, distance to villages and distance to rivers; rainfall, temperature, near-infrared and red reflectance, whereas topographic variables, except for elevation, did not contribute much to the current distribution of *Prosopis*. According to the random forest model, a total of 1.173 million ha (12.33% of the study region) was found to be invaded by *Prosopis* to varying degrees of cover. Our findings demonstrate that MLAs can be successfully used to develop fractional cover maps of plant species, particularly IAPS so as to design targeted and spatially explicit management strategies.

KEYWORDS

Afar Region, dryland ecosystems, Ethiopia, fractional cover mapping, invasive alien plant species, machine learning algorithms, *Prosopis juliflora*

1 | INTRODUCTION

In the last 20 years, many studies have attempted to accurately detect the spatial extent of invasive alien plant species (IAPS) to map their spread over time or model their potential invasion area. They have used a variety of environmental, bioclimatic, and/or earth observation data, and applying classification or regression methods. More recently, machine learning algorithms (MLAs) have gained high popularity in ecology and earth science because of their ability to model highly dimensional and non-linear data with complex interactions and deal with data gaps (Thessen, 2016). Good performances of MLAs have been obtained in several fields, including remote sensing classifications (Mountrakis, Im, & Ogole, 2011) and species distribution modeling (Cutler et al., 2007; Elith & Leathwick, 2009). However, for quantifying the impact of IAPS and developing spatially explicit management strategies, accurate information is crucial not only on the current or projected distribution of IAPS but also on their cover across the invaded range (Le Maitre, Gush, & Dzikiti, 2015; Shackleton, Le Maitre, van Wilgen, & Richardson, 2015a; Shackleton, Le Maitre, van Wilgen, & Richardson, 2015b). A few studies attempted to estimate fractional IAPS cover using remotely sensed data either applying spectral unmixing techniques (Frazier & Wang, 2011; Vilà et al., 2011) or using very high-resolution remotely sensed data, mostly in combination with machine learning classifiers (Cho, Malahlela, & Ramoelo, 2015; Masocha & Skidmore, 2011). The use of coarser resolution remote sensing resulted in accurate binary maps of presence and absence of IAPS (Chen, Yi, Qin, & Wang, 2017; Wakie, Evangelista, Jarnevich, & Laituri, 2014). Only recently, more promising mapping of IAPS at finer fractions of cover was obtained using a combination of medium or high-resolution satellite data and powerful machine learning classification algorithms (Ng et al., 2016; Rembold, Leonardi, Ng, Gadain, & Meroni, 2015). Such fine-scaled and accurate quantification of the local fractional cover of IAPS allows understanding their impacts through cover-impact curve analysis. Furthermore, it allows to identify areas with early stages of invasion where the control of satellite populations maybe halted or at least slow down further spread of IAPS (Vilà et al., 2011).

Prosopis juliflora (Swartz DC.), hereafter referred to as *Prosopis*, has been introduced to different parts of the world with the aim of providing benefits to rural people, such as the production of fuelwood, charcoal, or construction material (Engda, 2009; Haji & Mohammed, 2013; Mureriwa, Adam, Sahu, & Tesfamichael, 2016; Pasiecznik & Henry Doubleday Research Association, 2001). Like numerous other introduced plants, *Prosopis* has become invasive in many places and is increasingly known for its negative ecological and socio-economic impacts (Shackleton, Le Maitre, van Wilgen et al., 2015a; Shackleton, Le Maitre, van Wilgen et al., 2015b; van Wilgen & Wannenburgh, 2016). In Ethiopia, several studies have attempted to assess *Prosopis* distribution particularly in the Afar Region (Ayanu et al., 2014; Engda, 2009; Wakie et al., 2014), but they either focused on relatively small study areas or provided only coarse-resolution maps of either presence or absence of the species. Yet, at the early stage of its invasion, or at the invasion front, *Prosopis* often occurs in a patchy mixture with natural vegetation or as single trees, which is challenging to capture by remotely sensed data of moderate spatial resolution. Hence, the development of effective management strategies to mitigate the negative impacts of *Prosopis* requires accurate and detailed information on both invaded areas and on the level of invasion across the invaded area.

We set out to compare the performances of five MLAs (gradient boosting machine implemented in two different ways, random forest, support vector machine, and deep learning neural network), an ensemble model and a generalized linear model. This analysis helps identifying the best-performing algorithm in mapping detailed fractional cover of *Prosopis* in the dryland ecosystem of the Afar Region, Ethiopia. All model outputs were validated using a number of performance measures. The best-performing model was then used to create a *Prosopis* distribution and fractional cover map.

2 | METHODS

2.1 | Study area and study species

The study was conducted in the Afar National Regional State of Ethiopia (hereafter referred to as the Afar Region). The study area extends from 39.7°E to 42.4°E and 8.8°N to 14.5°N, and is located in the Great Rift Valley of Eastern Africa and covers an area of 9.51 million ha (Figure 1a). Mean annual rainfall is about 560 mm; and the mean annual temperature is about 31°C (MOA, 1997). The biome can be described as semi-arid to arid. Its vegetation cover consists of patches of scattered dry shrubs, acacia woodland (comprising different *Vachellia* species), bushland, grassland, and wooded grassland. People's main sources of livelihood are pastoralism and some agro-pastoralism around small rural towns (Yirgalem, 2001).

The study focuses on *Prosopis* species. *Prosopis* shows a wide range of ecological adaptations (from arid to tropical climate conditions) and occur along a large variety of environmental gradients (Asfaw & Thulin, 1989; Mohamed, 1997), including different soil types (from sand to heavy clays and stony soils) and a wide range of altitudes (from sea level up to 1,600 m. a.s.l: Shiferaw et al., 2019). Furthermore, *Prosopis* trees are able to fix nitrogen and have deep root systems, rendering them resistant to droughts (Keller, Lodge, Lewis, & Shogren, 2009; Mohamed, 1997). This has enabled *Prosopis* to become one of the most successful invasive woody plant species in arid and semi-arid areas. *Prosopis* has been planted to reclaim degraded land, combat desertification, reduce soil erosion (Mishra, Crews, & Okin, 2014; Pasiecznik & Henry Doubleday Research Association, 2001; Tessema, 2012; Wakie, Evangelista, & Laituri, 2012), and manage soil salinity (El-Keblawy & Al-Rawai, 2007). *Prosopis* trees originally planted in Ethiopia (Figure 1a) belong to the species *P. juliflora* (Figure 1b) in the late 1970s and early 1980s with the main aim of soil and water conservation (Pasiecznik & Henry Doubleday Research Association, 2001). However, since the early 1990s, its invasive nature has caused major problems in rangelands,

agricultural fields, and riverbanks, and aggravating conflicts on grazing land among pastoralists (Argaw, 2015; Kebede & Coppock, 2015; Tegegn, 2008). Such conflicts have been common in the Awash Basin, where *Prosopis* has invaded vast areas of precious rangeland and cropland (Wakie et al., 2012).

2.2 | Sampling design and datasets

Georeferenced field samples were collected throughout the entire study area using a stratified random sampling approach. Presence and absence plots were selected from invaded and uninvaded areas, respectively. Invaded areas were additionally stratified into heavily invaded and less invaded areas. Within those strata, careful attention was paid to collect representative samples of the entire cover gradient (0%–100%) of *Prosopis* coverage. In order to reduce spatial autocorrelation, each sampling plot had a minimum distance of 500 m to the next one. A total of 2,722 samples (presence and absence plots of 20 m × 20 m) were collected between September 2016 and March 2017. A plot was considered a presence plot if it contained at least one *Prosopis* plant; otherwise, it was considered an absence plot. About 70% of the samples were absence plots while 30% were presence plots. These shares were chosen based on a preliminary rough estimation of the shares of uninvaded and invaded land in the study area, which would avoid any bias of results toward either presence or absence of *Prosopis* (Jiménez-Valverde & Lobo, 2007). Finally, 80% of all sampling plots were randomly selected to be used for model calibration, whereas the other 20% were used for validation (Elith et al., 2011).

The spatial datasets were gathered from various sources and used as explanatory variables to run the models (Table 1). Explanatory variables differed in terms of spatial resolution, projection, and time of acquisition; thus, reprojection to UTM projection and nearest neighbor spatial resampling to a pixel resolution of 15 m was applied using panchromatic band of Landsat 8. The Landsat 8 (operational land imager-OLI) satellite data were acquired on 26 and 28 January as well as 11 and 20 February 2017 (paths: 167 and 168; rows: 50–54). In total, nine scenes were required to cover the entire study area and then mosaicked. These acquisition dates match the period of field data collection and fall into the study area's dry season, when herbs and grasses are dry and most trees and bushes except *Prosopis* have shed their leaves.

The remotely sensed datasets were checked for geometric correspondence to all other datasets. Further, these datasets were atmospherically corrected using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm (Chavez, 1996; Lu, Mausel, Brondizio, & Moran, 2002). The Red, the near-infrared (NIR) and the first shortwave-infrared (SWIR1) bands of Landsat 8 were selected as explanatory variables. Furthermore,

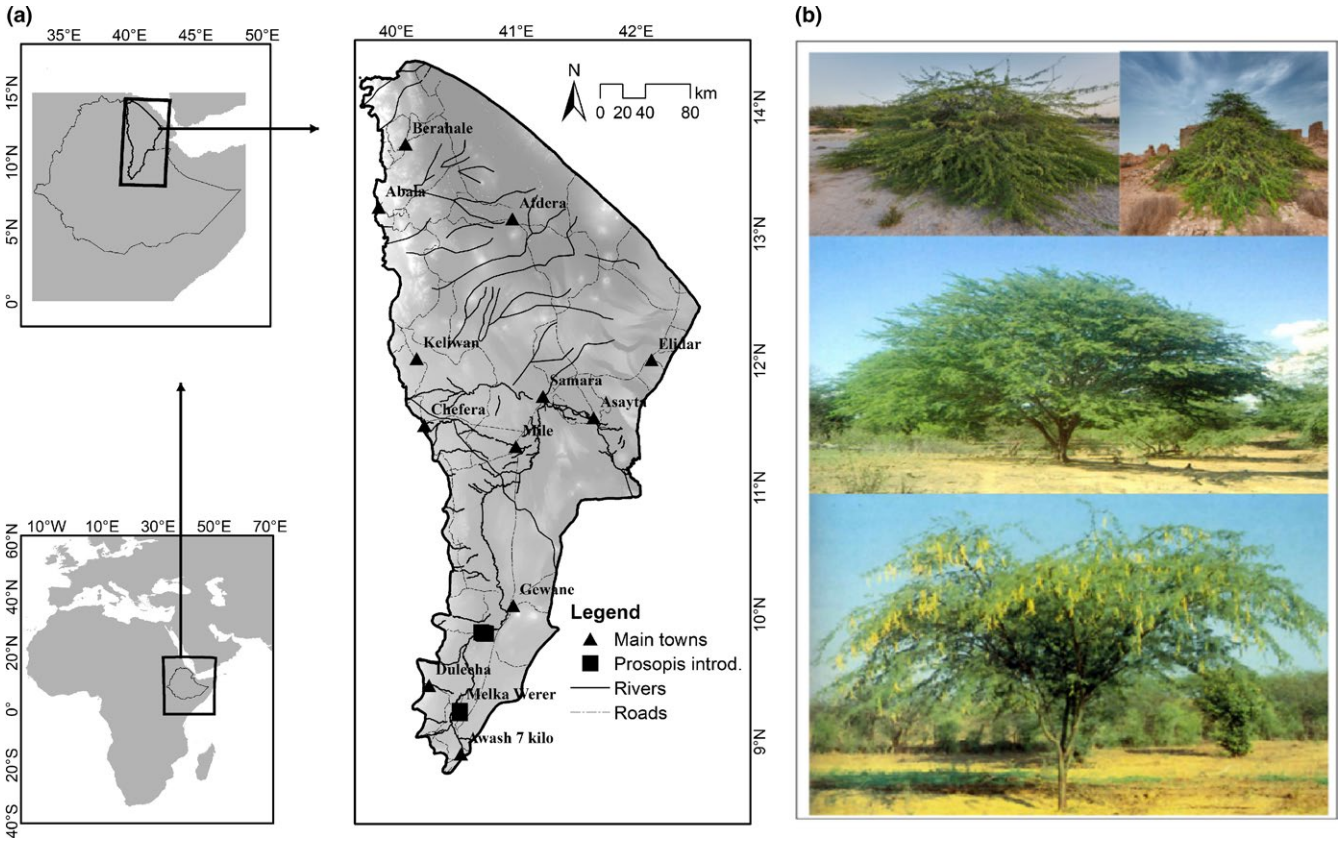

FIGURE 1 Location of the study area, Afar National Regional State, in Ethiopia (a). The detailed map shows the main towns, roads, and rivers, as well as the locations where *Prosopis* was first introduced. The shading indicates elevation, ranging from 175 m below sea level (dark gray) to 2,992 m above sea level (white), and photos of *Prosopis* plant (b)

the normalized difference vegetation index (NDVI) was calculated from Red and NIR bands and used as another input. All selected bands, as well as the NDVI, have proven to be particularly suitable to capture photosynthetic active vegetation, and soil and vegetation moisture content (Barsi, Lee, Kvaran, Markham, & Pedelty, 2014). Additionally, daytime (LSTd) and nighttime (LSTn) land surface temperatures from MODIS sensor were included. In order to have longer-term LSTd and LSTn average data, a 5-year average of these products was generated between 2012 and 2017. The spatial resolution of these datasets is 1 km. Although this seems rather low compared to the other datasets, these day- and nighttime temperature datasets have shown to be useful in species distribution modeling and, particularly in Africa where weather stations are scarce. These datasets have shown to be more accurate than other global climate datasets (Ashby, Moreno-Madriñán, Yiannoutsos, & Stanforth, 2017; He et al., 2015). Moreover, we used variables representing topography, infrastructure as well as watercourses as these variables have shown to have an influence on *Prosopis* distribution (Shiferaw et al., 2019).

2.3 | Models

Our study evaluates the performances of seven algorithms in mapping *Prosopis* distribution and fractional cover abundance. We chose five MLAs: two different implementations of gradient boosting machine (GBM and GBM-BRT), random forest (RF), support vector machine (SVM), and deep (learning) neural network (DNN), an ensemble model composed of the four best-performing tested algorithms, and a generalized linear model (GLM) for comparison reasons. All model calculations and model performance assessments were implemented in R programming (R Core Team, 2017). A comprehensive overview of the R packages and the different parameter settings are provided in the Supporting information (Table S1). We checked collinearity of explanatory variables before applying to any model and those having high variable inflated factors (VIF) were removed. In this study, we used a threshold level of VIF > 10 to exclude variable(s) from any model (Bruce & Bruce, 2017; Gareth, Witten, Hastie, & Tibshirani, 2014). Accordingly, three variables, the blue, green and second shortwave-infrared (SWIR 2) bands were removed from all models as they had high VIF (Dormann et al., 2012). We then assessed the influence (importance) of variables in each model by using the method described by Natekin and Knoll (2013). Furthermore, 10-fold cross-validation was applied to assess model performance (Fushiki, 2009). Finally, the predictive power of all tested MLAs was evaluated using several performance parameters (Table 2). The general functionality of each tested model is described below.

Until few years ago, multivariate linear regression was the most commonly used approach in species distribution modeling (Collingham, Wadsworth, Huntley, & Hulme, 2000; Higgins et al., 2003; Stohlgren et al., 2010). In this study, the GLM was included to compare the performance with the MLAs (Nicholls, 1989; Getis &

TABLE 1 List of spatial data and explanatory variables used for the modeling of *Prosopis* fractional cover

Variable abbreviations	Description	Source
Rain	Mean annual rainfall	Ethiopian National Meteorol. Agency
Temp	Mean monthly temperature	
LSTd	Monthly land surface temperature during daytime and nighttime; for the modeling 5-year averages were calculated	MODIS, NASA
LSTn	Monthly land surface temperature during nighttime; for the modeling 5-year averages were calculated	MODIS, NASA
PAN	Panchromatic reflectance	Landsat 8 OLI, USGS
Red	Red reflectance	Landsat 8 OLI, USGS
NIR	Near-infrared reflectance	Landsat 8 OLI, USGS
SWIR1	Shortwave-infrared band 6 reflectance	Landsat 8 OLI, USGS
NDVI	Normalized difference vegetation index	
Elevation	Shuttle Radar Topography Mission digital elevation model (30 m spatial resolution)	USGS
Slope	Derived from elevation	
Relief	Derived from elevation (contour) differences	Adediran, Parcharidis, Poscolieric, and Pavlopoulos (2004)
Landform	Topographic position index derived from elevation, aspect and slope	Dikau (1989); Dikau, Brabb, & Mark (1991); Weiss (2001); Ilia, Rozos, & Koumantakis (2013)
Rugged	An index derived from elevation	Riley, DeGloria, & Elliot (1999)
DistRoad	Distances derived from road network data	Ethiopian Road Authority
DistVillage	Distances derived from settlement data	EthioGIS and Central Statistical Agency
DistRiver	Distances derived from data on watercourses	EthioGIS

Ord, 1992). We used backward and forward stepwise variable selection to find a parsimonious model (Pearce and Ferrier, 2000). Akaike Information Criterion was used as the model performance metric (step-AIC; Higgins et al., 2003).

Gradient boosting machine as well as GBM-BRT use a boosting approach where datasets are resampled several times to generate results that form a weighted average of the resampled dataset. This is done by creating a gradient (or step-by-step) boosting by minimizing errors among series of decision trees that together form a single predictive model (Natekin & Knoll, 2013; Olinsky, Kennedy, & Kennedy, 2012; Wana & Beierkuhnlein, 2010; Boser,Guyon, & Vapnik, 1992). In our study, we tested two implementations of GBM and GBM-BRT. They are both based on the same packages: "gbm," "caret," "dismo," and "raster," with "dismo" and "caret" using the "gbm" package to fit the models. The main differences of the two implementations are the use of different hyper-parameters. We varied the interaction depth (i.e., tree complexity in GBM-BRT) which we set to 3 for GBM and was set to 5 for GBM-BRT, as well as the loss function. While GBM used the "Gaussian" family (Friedman, 2001), GBM-BRT used the "Bernoulli" (Elith, Leathwick, & Hastie, 2008). Furthermore, the final selection of number of trees and the learning rate was different. We tuned the models by only varying the number of trees and the number of repeats while other parameters were kept stable using their respective R package default settings (for details see also Supporting information Table S1). Fine-tuning the number of iterations is done to improve the performance of a model by fitting either many sub-models or gradient fitting and combining them for final prediction. All models were tuned using the same performance metrics. For the fine-tuning, we calculated mean change in predictive deviance ±one standard error (Elith et al., 2011). The optimization of the number of trees improved the performance substantially (Supporting information Figure S1).

The RF builds the trees in parallel processes (Breiman, 2001). The trees are fully grown and each is used to predict the out-of-bag observations that do not occur in a bootstrap sample (Breiman, 2001). The predicted class of an out-of-bag observation is calculated average of the results of all predictions (Breiman, 2001; Youssef, Pourghasemi, Pourtaghi, & Al-Katheeri, 2016). The RF has some

TABLE 2 Parameters used to assess model performance

Perf ormance parameter	Description	Sources
Confidence interval (CI)	It provides a range of values within which the population parameter is likely to lie. In a normal distribution, the general expression of the confidence interval is: Estimate ± $\frac{Z\alpha}{2}(SE)$, where *SE* is the standard error of the estimate and, if $\alpha = 0.05$, $z = 1.96$. The provision of confidence limits in addition to accuracy is particularly useful in comparative analyses	Newcombe (1998)
Correlation	Agreement between fractional cover measured in the field samples and the predicted fractional cover for the same samples	Harrington (2006); Meynard & Quinn, (2007)
Sensitivity	Known as true-positive rate (TPR); measures the proportion of positives that were correctly identified as locations where *Prosopis* was present. Calculated as: $\frac{TP}{(TP+FN)}$; where TP stands for true positives, and FN for false negatives	Metz (1978); Fuchs, DeMeester and Albertucci (1987)
Specificity	Known as true-negative rate (TNR); measures the proportion of negatives that were correctly identified as locations where *Prosopis* was absent. Calculated as:$\frac{TN}{(TN+FP)}$;where TN stands for true negative, and FP for false positives	Fuchs et al. (1987)
Accuracy	Class accuracy is calculated by dividing the number of correct pixels in that category by the total number of pixels in either the corresponding row or the corresponding column; it indicates the probability of a reference pixel being correctly classified and is really a measure of omission error. Calculated as: $\frac{TP+TN}{(TP+FP+TN+FP)}$; where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives	Congalton (1991) Fuchs et al.(1987)
AUC	Area under the receiver operating characteristics (ROC) curve; indicates the model's accuracy in handling true values (presence of *Prosopis*) as true and false values (absence of *Prosopis*) as false. The higher the AUC, the better the model fit, and vice versa	Landis & Koch (1977); Metz (1978)
Kappa coefficient	Statistical measure of inter-rater agreement, excluding agreements occurring by chance. It is calculated in a confusion matrix as $\frac{(0.5\times TP)}{(TP+FN)} + \frac{(0.5\times TN)}{(TN+FP)}$	Metz (1978)
Balanced accuracy	Average of all class accuracies; takes into account unbalanced class sizes. In our case, with two classes (presence and absence of *Prosopis*) it is calculated as: $\frac{1}{2}\left(\frac{TP}{P}+\frac{TN}{N}\right)$	Brodersen, Ong, Stephan, and Buhmann (2010)
Threshold (max @ TPR + TNR)	Maximum value at which the true-positive rate (TPR, or sensitivity) and the true-negative rate (TNR, or specificity) intersect. It is often used as a threshold level in dichotomies. In our case, values above the threshold indicate that *Prosopis* is present; values below the threshold indicate that *Prosopis* is absent	Metz (1978); Getis & Ord (1992); Hijmans and Elith (2015)

limitations like incapable of predicting beyond the range of response values in the training data (Hengl et al., 2015), and overestimate low values and underestimate high values (Horning, 2010). In this study, we only varied the number of trees, testing two different settings: 1,000 and 5,000 trees while all other parameters were set to default.

The SVM can be used for classification or regression. It constructs a hyperplane or set of hyperplanes in an infinite-dimensional space and tries to find the optimal separating hyperplanes, that is, the planes where the separability between classes is at its maximum (Noble, 2006; Rodrigues & De la Riva, 2014). The SVMs have many mathematical features that make them attractive for prediction, handle extremely high-dimensional feature spaces, and identify outliers (Brown et al., 2000; Kimothi & Dasari, 2010). We varied settings for the kernel, the cost function and gamma (Supporting information Table S1).

The DNN has become very popular recently but is still sparsely used by the geoscience community (Zhang, Zhang, & Du, 2016). The DNN is fully connected neural networks composed of multiple hidden layers together with non-linear transformations and a variety of tailored architectures (Guo et al., 2016). The DNN has a capacity to analyze big data. In this study, we used a feed-forward neural network.

The purpose of ensemble models is that it should combine the benefits of each included optimized model and penalize the overestimate or underestimate of each individual model. Thus, in order to be able to do so they should be diverse and complement each other on the one hand, but also each one of them independently achieving a high performance (Chitra & Uma, 2010). Our ensemble model consisted of the four best-performing models (RF, GBM, SVM, and GLM). They were weighted using the function "glmnet" where the predictions from each model are used as a predictor in a GLM and the resulting GLM coefficients determine how much each model should be weighted (Hastie & Qian, 2016; R Core Team, 2017). The coefficients of contribution of each model in our ensemble were 0.2 for RF, 0.1 for GBM, 0.05 for SVM, and 0.01 for GLM as indicated in Figure 2.

3 | RESULTS

3.1 | Model parameter settings and weighting of variables

Optimum performance of the GBM-BRT was found when using ~6,050 trees than 3,100 trees; while the GBM performed better with 500 trees than 100 trees. The RF model performed better for 5,000 trees than with 1,000 trees. The tested algorithms weighted the explanatory variables differently, depending on each model's sensitivity to small variations in the data and to the variable types (Figure 2a–g). In all models, Relief, Landform, Rugged, and Slope were removed again from the model except from the DNN. In the DNN, the least important variables were NIR, PAN, Red, NDVI, and Slope. Interestingly, DistRoad, Rugged, Relief, and Slope proved to be among the least-contributing variables in the GLM model and were removed from the final iteration (Figure 2a) though DistRoad was one of the important contributors in other models. In the MLAs, 13 out of the 17 variables were kept. The most important explanatory variables, having >5% relative influence, were selected by more than one MLAs. These are NDVI, Elevation, DistVillage, DistRiver, Rain, NIR, Red, LSTd, and LSTn in decreasing order. The first four variables had the highest influence in four of the seven models to explain *Prosopis* distribution (Figure 2).

3.2 | Evaluation of the models

Among the tested models, the RF performed the best, followed by the ensemble model, GBM and SVM (Table 3). The last two performed comparably. While the GBM achieved slightly higher accuracies and kappa statistics than the SVM, but the SVM obtained better sensitivity and specificity scores. While the GBM-BRT achieved high accuracy compared to the GLM but its kappa, sensitivity, and specificity scores are low. However, the GLM's specificity score was higher than the ones obtained by the GBM-BRT model. DNN did not perform well. Its sensitivity and specificity scores were very unbalanced and its sensitivity score was very low. All models performed better in terms of specificity than sensitivity. This indicates that uninvaded areas (true absence rate) were better identified and classified than invaded areas (true presence rate).

3.3 | *Prosopis* fractional cover

Comparing the results of different models, we found considerable variation in the extent of invaded areas, even though we used the same input datasets for all algorithms. The most extreme estimates of the total area invaded by *Prosopis* were generated; the highest was from the DNN model (34.8% invaded) and the lowest was from the SVM model (11.2% invaded). The best-performing RF model, calculated the total invaded area to be 12.33% of the Afar Region. The other four models—the GBM-BRT, the ensemble model, the GBM, and the GLM estimated the total invaded area by *Prosopis* at 16.1%, 14.9%, 14.7%, and 20.1%, respectively (Figure 3 and Table 3). Hence, the results produced by the ensemble model, the SVM, and the GBM were fairly close to that produced by the RF model (Table 3).

Following the evaluation of the different models, the best-performing RF model was used to map the current fractional cover of *Prosopis* in the Afar Region. The RF model's sensitivity and specificity values suggest that the model is robust, and its AUC value indicates that the presence of *Prosopis* was correctly mapped with a probability of 97%. A threshold value of 0.326 was calculated from the model for the minimum cover level of *Prosopis* presence, which corresponds to 0.4% *Prosopis* fractional cover found on the ground (Figure 4). According to the RF prediction, about 1.173 million ha of land is invaded by *Prosopis* at different stages of cover abundances in the Afar Region.

4 | DISCUSSION

4.1 | Model optimization

During model optimization, the number of trees (for the GBM-BRT, GBM, and RF), the learning rate (sets the weight applied to

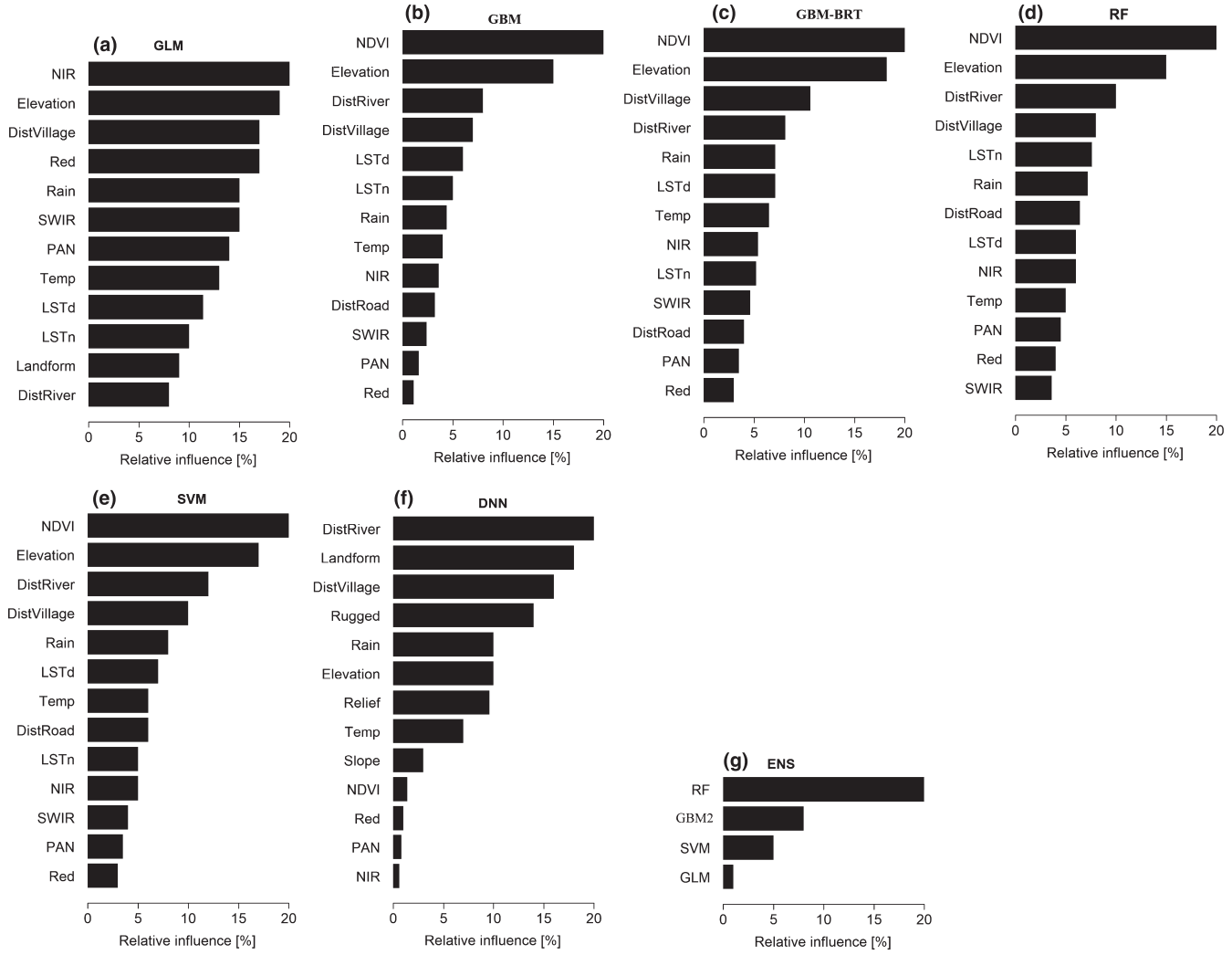

FIGURE 2 Relative influence of explanatory variables in the different algorithms after removal of the least-contributing ones: (a) generalized linear model (GLM), (b) gradient boosting machine (GBM), (c) gradient boosting machine using boosted regression trees package (GBM-BRT), (d) random forest (RF), (e) support vector machine (SVM), (f) deep learning neural network (DNN), (g) ensemble model (ENS)

individual trees), and the bag fraction (which sets the proportion of observations) had the greatest influence on model performance (Elith & Leathwick, 2009). For example, the lower of two learning rates tested in the GBM-BRT required more trees, which improved the result without causing overfitting (Mining, 2009; Hijmans & Elith, 2013, 2015). Consequently, the lower learning rate of 0.005 with 6,050 trees performed better than that of 0.01 with 3,100 trees. However, a learning rate of 0.0025 with 10,000 trees did not perform better than that of 0.005 in the GBM-BRT even though the increase in the number of trees reduced deviance, eventually stabilizing the model. This indicates that lowering the learning rate without comparing model performance would have resulted in two disadvantages: a poorer model fit and longer computational time without improving the model's accuracy.

Variable reduction contributed to model stability, which is evident in the GBM-BRT and GLM models. Similar studies in the GBM showed model stability after variable reduction (Getis & Ord 1992; Burnham & Anderson, 2002). Removal of the topographic variables such as Rugged, Landform, Relief, and Slope from most of the tested models indicates that these variables contributed little to the models' performances. Also, except through Elevation, topography does not seem to add significant information regarding the current distribution and cover of *Prosopis* in the study area. This is probably because the study area is largely flat. The DNN model produced one of the least accurate results. The ROCs of GLM and DNN showed different from other MLAs (Supporting information Figure S2). The ROC curve in the GLM nears quickly the 100% true positives rate but the ROC curve in the DNN remains flat achieving a comparably high amount of false-positive rate compared to its true-positive rate. Different reasons could have led to a poor performance, for example, batch size may be small to the DNN. But then there are things like to check for hidden dimension layers, analyze the gradient checks. Further tuning might have been necessary to improve the DNN (change a different optimizer, change regularization, check and adjust weights at initialization, etc). However, this requires further investigation.

TABLE 3 Summarized performance parameters of the evaluation of current fractional cover maps of *Prosopis* in the Afar Region produced by means of different models. Additionally, AUC plots for each model are provided in the Supporting information Figure S2

Model type	95% CI	Accuracy	Kappa	Balanced accuracy	Sensitivity	Specificity	Pos. pred. value	Neg. pred. value	AUC	Correlation	Threshold
GLM	0.763, 0.834	0.801	0.498	0.744	0.612	0.882	0.651	0.858	0.852	0.564	0.285
GBM	0.837, 0.897	0.877	0.678	0.848	0.802	0.895	0.738	0.924	0.944	0.756	0.397
GBM-BRT	0.761, 0.841	0.789	0.316	0.727	0.632	0.712	0.728	0.868	0.945	0.794	0.258
RF	0.897, 0.945	0.918	0.797	0.911	0.894	0.926	0.818	0.959	0.971	0.829	0.326
SVM	0.856, 0.907	0.872	0.677	0.827	0.817	0.918	0.864	0.891	0.876	0.741	0.151
DNN	0.689, 0.767	0.729	0.434	0.574	0.014	0.995	0.568	0.724	0.595	0.206	0.392
Ensemble	0.871, 0.925	0.891	0.771	0.873	0.846	0.919	0.808	0.939	0.962	0.841	0.349

Note. DNN: deep learning neural network; ENS or ensemble: ensemble model; GBM: gradient boosting machine; GBM-BRT: gradient boosting machine using boosted regression trees package; GLM: Generalized linear model; RF: random forest; SVM: support vector machine.

4.2 | Important variables

Among the infrastructure variables, DistVillage was found to be important in all models except in the GLM. Among the environmental variables, Elevation was the most important explanatory variable for the distribution and fractional cover of *Prosopis* in all models except the DNN. While NDVI and DistRiver had a high relative importance in the MLAs, they were removed from the GLM during variable reduction. From a methodological perspective, this suggests that the GLM is not able to relate variables having a linear or radial spatial pattern to the samples used in the models, and therefore, is less suited to explain *Prosopis* distribution and fractional cover. It is well known that *Prosopis* is primarily spread by livestock (Shiferaw, Teketay, Nemomissa, & Assefa, 2004), human transport and along watercourses, thereby promoting discontinuity or jump dispersal (Wilson, Dormontt, Prentis, Lowe, & Richardson, 2009). However, the GLM was not able to fully capture these phenomena. In the DNN model, Landform exceptionally ranked second in importance, following DistRiver.

The influences of the tested explanatory variables varied in terms of magnitude and direction depending on each model's sensitivity. In the case of NDVI, this is in line with the general observation of greenness, and therefore, also NDVI, increases with increasing *Prosopis* cover. It suggests that particularly NDVI captured during dry season is a good variable for explaining the current distribution of *Prosopis* due to the plant's evergreen behavior in the study area unlike other plant species shed their leaves during the dry season. The explanatory power of NDVI is further supported by the fact that greenness or NDVI is a consequence of *Prosopis* presence and cover level but not a cause of its distribution. Our results also show that *Prosopis* cover increases with increasing temperature. *Prosopis* grows best in arid and semi-arid environments and can stand air temperatures of up to 50°C (Mohamed, 1997). Besides temperature, elevation had a strong influence on *Prosopis* distribution in the study area as *Prosopis* cover increases with decreasing elevation.

As mentioned above the main causes of dispersal are by livestock, human transport and by water which explains well the strong influence of these factors in most models. In contrast to Menuz & Kettenring (2013), our data suggest that landscape structure variables are more relevant for species distribution/invasion at the current stage of invasion than climatic factors (precipitation and temperature), which describe the environmental niche of plant species (Guisan & Thuiller, 2005). However, at larger spatial scales climatic factors might additionally capture well the distribution pattern of the species (Coutts, Klinkenvan, Yokomizo, & Buckley, 2011; Cabra-Rivas, Saldana, Castro-Dıez, & Gallien, 2016).

4.3 | Fractional cover of *Prosopis*

Different algorithms produced different results with varying accuracies. Thus, these algorithms differ in their sensitivity (power to distinguish *Prosopis* distribution from other vegetation) across spatial variabilities. In this study, we found the RF to be the best-performing

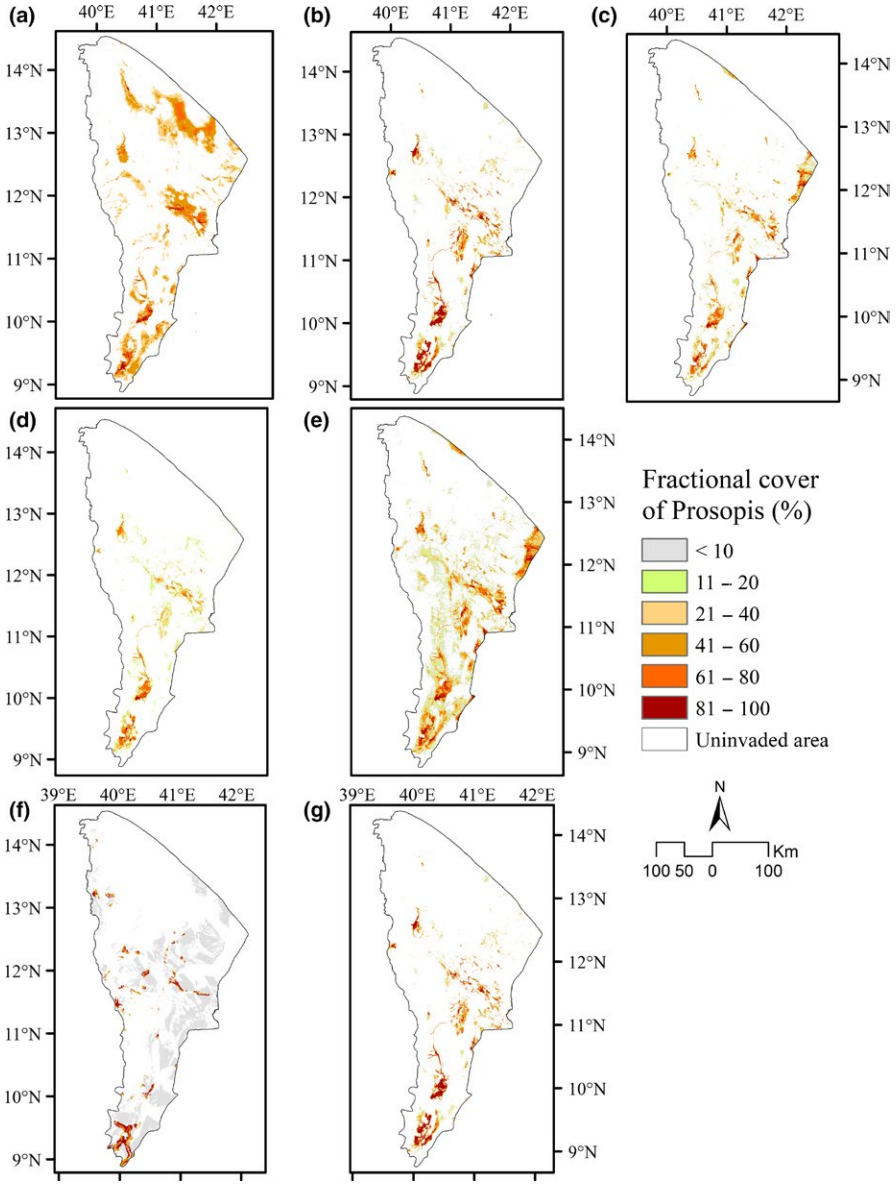

FIGURE 3 The current fractional cover maps of *Prosopis* distribution were produced by using different machine learning algorithms. (a) generalized linear model (GLM), (b) gradient boosting machine (GBM), (c) gradient boosting machine using boosted regression trees package (GBM-BRT), (d) random forest (RF), (e) support vector machine (SVM), (f) deep learning neural network (DNN), (g) ensemble model (ENS)

algorithm (AUC = 0.971, κ = 0.797). Surprisingly, the ensemble model (AUC = 0.962, and κ = 0.771) performed slightly less than the RF, although other studies had suggested that an ensemble model would be able to overcome some of the individual models' limitations (Kim, 2017) and expected to obtain better performance. Our finding indicates that some of the models included in the ensemble model might have introduced errors, thereby impairing or penalizing its performance. Also, the DNN did not perform well which we cannot fully explain. Reasons could be the DNN may not be appropriate for species distribution mapping and requires further investigations.

Application of a threshold level to produce binary maps of presence and absence has been tested (Zhou, Chen, Cao, & Chen, 2015). In this study, we also applied threshold levels to distinguish invaded from uninvaded areas with a threshold level of the RF model at 0.326. Based on this threshold, we found a very large area (~1.173 million ha) to be invaded. Our result is in line with the amount of invaded areas estimated by MoLF (2017) to be about 1.2 million ha.

Detection of the spread and establishment of an invasive plant species is highly important for an effective management at an early stage of invasion (i.e., low to medium cover levels). Soft classification, as performed in this study, based on satellite data, climatic, topographic, and other relevant data enables not only identification of a particular species but also retrieval of that species' fractional cover even at low cover fractions. Another interesting finding is that all models performed better in terms of specificity than sensitivity (Table 3). This indicates that uninvaded areas (true absence rate) were better identified and classified than invaded areas (true presence rate). A reason for this may be that the model sometimes misinterpreted acacia shrubs present in invaded areas as *Prosopis*; otherwise the unbalanced of sample size between presence and absence doesn't affect the quality the output (Jiménez-Valverde & Lobo, 2006) as long as enough sample size were used from each group.

Machine learning algorithms have attracted significant attention in the modeling community. First, shallow Neural Networks

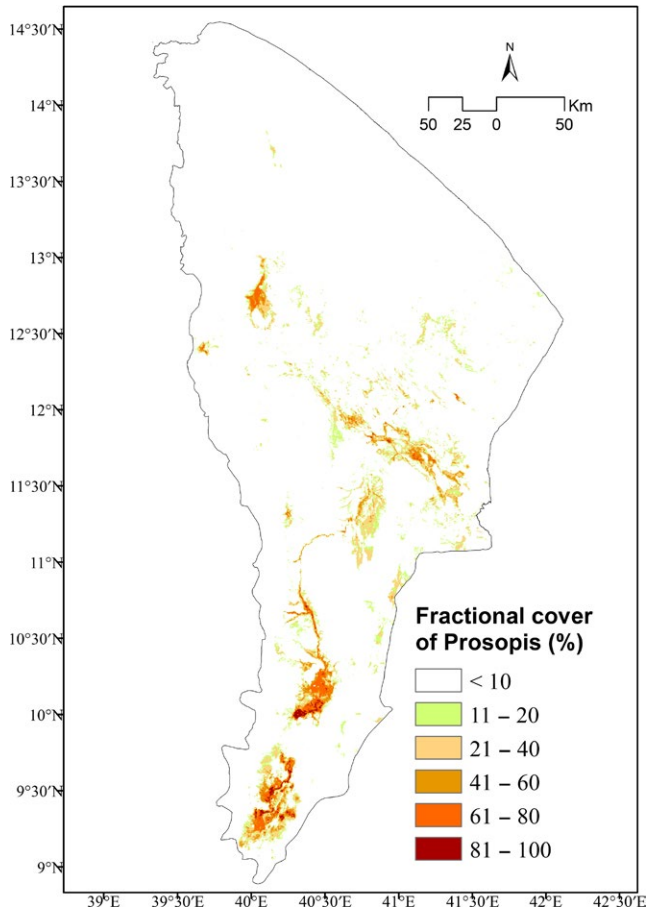

FIGURE 4 Current fractional cover of *Prosopis* (after matching to the ground cover level) in the Afar Region according to the RF model. For better readability, fractional cover was grouped in six fractional cover classes

(NN) attracted a lot of attention and were widely applied to many different research problems (Zhou et al., 2015). In the remote sensing community, the DNN was soon followed by other MLAs: the GBM-BRT, the SVM, and the RF, which provided better results both in regression and classification (Ashby et al., 2017; Pal & Mather, 2005, 2017). Our regression analyses in the present study indicated that the RF, the ensemble model, and the GBM outperformed the SVM. A similar finding was reported by Lorena et al. (2011), who compared the performances of the RF and the SVM in modeling the potential distribution of 35 species in Brazil. The study by Mi, Huettmann, Guo, Han, and Wen (2017) also indicated that the RF performed better than other algorithms tested to model crane species.

Our finding confirms that the RF is a suitable algorithm for fractional cover mapping of plant species. However, based on our experiences gained during this study five important points should be considered in order to achieve good results while applying the RF regression: (a) sufficient and well-distributed field data samples should be collected in the study area; (b) the number of presence and absence field samples should be proportional to the shares of the study area where the species is present and absent, respectively; (c) the field data values for the dependent variable should be within the range of the expected prediction values, (d) as shown by previous study, the values of explanatory variables used for training need to represent the entire range of values present in the study area (Hengl et al., 2015), and (e) fine-tuning of algorithm parameters and variable reduction are recommended for improved model fitness and better regression outputs.

5 | CONCLUSIONS

Fine-scaled fractional cover maps of IAPS are a key requisite for estimating the environmental and socio-economic impacts of IAPS and for designing spatially explicit management strategies. Our findings show that the RF regression is outperformed other algorithms and is a suitable for mapping the fractional cover of species distribution in agro-climatic contexts similar to those of the Afar Region. While the GBM and the SVM achieved only slightly less accurate results, the GLM, the GBM-BRT, and the DNN did not perform well when looking at sensitivity, specificity, kappa, and the AUC. Nevertheless, performances of MLAs might be different if a much larger amount of data (i.e., predictor variables) is used, or if less training data is available or if the study is done in a different agro-ecological context. For this reason, we recommend evaluating the performances of two or more algorithms regarding the specific tasks required and the specific environmental settings prevailing in the context of plant species distributions.

ACKNOWLEDGMENTS

We are grateful for financial support from the Swiss Programme for Research on Global Issues for Development (r4d), funded by the Swiss National Science Foundation (SNSF) and the Swiss Agency for Development and Cooperation (SDC), for the project "Woody invasive alien species in East Africa: Assessing and mitigating their negative impact on ecosystem services and rural livelihood" (Grant Number: 400440_152085). We thank the staff of Addis Ababa University's Water and Land Resource Center for their support and facilitation services, Marlène Thibault from University of Bern Centre for Development and Environment (CDE) for critically editing the English language, and Urs Schaffner of CABI-Switzerland for the critical comments of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTION

HS designed the sampling procedure, carried out the survey, and performed the modeling and the calculations; WB and SE provided technical advice; and all authors contributed to the writing of the manuscript.

DATA AVAILABILITY

The dataset will be archived in an appropriate public archive.

ORCID

Hailu Shiferaw https://orcid.org/0000-0002-3697-083X
Sandra Eckert https://orcid.org/0000-0002-9579-5680

REFERENCES

Adediran, A., Parcharidis, I., Poscolieric, M., & Pavlopoulos, K. (2004). Computer-assisted discrimination of morphological units on north-central Crete (Greece) by applying multivariate statistics to local relief gradients. *Geomorphology*, *58*, 357–370.

Asfaw, H., & Thulin, M. (1989). Mimosoideae. In: I. Hedberg. & S. Edwards (Eds.), *Flora of Ethiopia* (pp. 71–73). Uppsala: National Herbarium, Addis Ababa University Addis Ababa, Uppsala University.

Argaw, T. (2015). Impacts of utilizing invasive *Prosopis juliflora* (Swart) DC on rural household economy at Gewane District, Afar Regional. *Journal of Economics and Sustainable Development*, *6*(5), 81–98.

Ashby, J., Moreno-Madriñán, M., Yiannoutsos, C., & Stanforth, A. (2017). Niche modeling of dengue fever using remotely sensed environmental factors and boosted regression trees. *Remote Sensing*, *9*(4), 328. https://doi.org/10.3390/rs9040328

Ayanu, Y., Jentsch, A., Müller-Mahn, D., Rettberg, S., Romankiewicz, C., & Koellner, T. (2014). Ecosystem engineer unleashed: *Prosopis juliflora* threatening ecosystem services? *Regional Environmental Change*, *15*(1), 155–167. https://doi.org/10.1007/s10113-014-0616-x

Barsi, J. A., Lee, K., Kvaran, G., Markham, B. L., & Pedelty, J. A. (2014). The spectral response of the Landsat-8 operational land imager. *Remote Sensing*, *6*(10), 10232–10251. https://doi.org/10.3390/rs61010232

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *In 5th Annual ACM Workshop on COLT* (pp. 144–152). Pittsburgh, PA: ACM Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brodersen, K., Ong, C., Stephan, K., & Buhmann, J. (2010). The balanced accuracy and its posterior distribution. *International Conference on Pattern Recognition*, 3122–3124.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., … Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(1), 262–267. https://doi.org/10.1073/pnas.97.1.262

Bruce, P., & Bruce, A. (2017). *Practical statistics for data scientists*. Sebastopol, CA: O'Reilly Media.

Burnham, K. P., & Anderson, D. R.. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (Vol. *172*, 2nd ed.). New York, NY: Springer. Ecological Modelling. https://doi.org/10.1016/j.ecolmodel.2003.11.004.

Cabra-Rivas, I., Saldana, A., Castro-Dıez, P., & Gallien, L. (2016). A multiscale approach to identify invasion drivers and invaders' future dynamics. *Biological Invasions*, *18*, 411–426. https://doi.org/10.1007/s10530-015-1015-z

Chavez, P. S. (1996). Image-based atmospheric corrections – Revisited and improved. *Photogrammetric Engineering and Remote Sensing*, *62*(9), 1025–1036.

Chen, J., Yi, S., Qin, Y., & Wang, X. (2017). Improving estimates of fractional vegetation cover based on UAV in alpine grassland on the Qinghai – Tibetan Plateau Improving estimates of fractional vegetation cover based on. *International Journal of Remote Sensing*, *37*(8), 1922–1936. https://doi.org/10.1080/01431161.2016.1165884

Chitra, A., & Uma, S. (2010). An ensemble model of multiple classifiers for time series prediction. *International Journal of Computer Theory and Engineering*, *2*(3), 454–458. https://doi.org/10.7763/IJCTE.2010.V2.184

Cho, M. A., Malahlela, O., & Ramoelo, A. (2015). Assessing the utility WorldView-2 imagery for tree species mapping in South African subtropical humid forest and the conservation implications: Dukuduku forest patch as case study. *International Journal of Applied Earth Observation and Geoinformation*, *38*, 349–357. https://doi.org/10.1016/j.jag.2015.01.015

Collingham, Y. C., Wadsworth, R. A., Huntley, B., & Hulme, P. E. (2000). Predicting the spatial distribution of non-indigenous riparian weeds: Issues of spatial scale and extent. *Journal of Applied Ecology*, *37*(Suppl 1), 13–27. https://doi.org/10.1046/j.1365-2664.2000.00556.x

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing Environment*, *37*, 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Coutts, S., Klinkenvan, V., Yokomizo, H., & Buckley, Y. (2011). What are the key drivers of spread in invasive plants: Dispersal, demography or landscape: And how can we use this knowledge to aid management? *Biological Invasions*, *13*, 1649–1661. https://doi.org/10.1007/s10530-010-9922-5

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Dikau, R. (1989). The application of a digital relief model to landform analysis. In J. F. Raper (Ed.), *Three dimensional applications in geographical information systems* (pp. 51–77). London, UK: Taylor and Francis.

Dikau, R., Brabb, E., & Mark, R. (1991). *Landform classification of New Mexico by Computer. U.S. Geological Survey Open File Report*, 91–634.

Dormann, C., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2012). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *35*, 001–020. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Leathwick, R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *2008*(77), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists: Statistical explanation of MaxEnt. *Diversity and Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

El-Keblawy, A., & Al-Rawai, A. (2007). Impacts of the invasive exotic *Prosopis juliflora* (Sw.) D.C. on the native flora and soils of the UAE. *Plant Ecology*, *190*(1), 23–35. https://doi.org/10.1007/s11258-006-9188-2

Engda, G. (2009). *Spatial and Temporal Analysis of Prosopis juliflora (Swarz) DC Invasion in Amibara Woreda of the Afar NRS*. MSc Thesis, AAU. Retrieved from http://etd.aau.edu.et/handle/123456789/786

Fuchs, K., DeMeester, T., & Albertucci, M. (1987). Specificity and sensitivity of objective diagnosis of gastroesophageal reflux diseases. *Surgery*, *102*(4), 575–580.

Frazier, A. E., & Wang, L. (2011). Characterizing spatial patterns of invasive species using sub-pixel classifications. *Remote Sensing of Environment*, *115*(8), 1997–2007. https://doi.org/10.1016/j.rse.2011.04.002

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Fushiki, T. (2009). *Estimation of Prediction Error by Using K-fold cross-validation*, https://doi.org/10.1007/s11222-009-9153-8)

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in R*. New York, NY: Springer Publishing Company Incorporated.

Getis, A., & Ord, J. (1992). The analysis of spatial association by use of distance statistics in geographical analysis. *Geographical Analysis*, *24*(3), 189–206.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48. https://doi.org/10.1016/j.neucom.2015.09.116

Haji, J., & Mohammed, A. (2013). Economic impact of *Prosopis juliflora* on agropastoral households of Dire Dawa administration, Ethiopia. *African Journal of Agricultural*, *8*(9), 768–779. https://doi.org/10.5897/AJAR12.014

Harrington, P. (2006). Statistical validation of classification and calibration models using bootstrapped Latin partitions. *TrAC Trends in Analytical Chemistry*, *25*(11), 1112–1124.

Hastie, T., & Qian, J. (2016). *Glmnet Vignette*, Stanford.

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M. N., Schmidtlein, S., … Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, *1*(1), 4–18. https://doi.org/10.1002/rse2.7

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., … Tondoh, J. E. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE*, *10*(6), 1–17. https://doi.org/10.1371/journal.pone.0125814

Higgins, S. I., Clark, J. S., Nathan, R., Hovestadt, T., Schurr, F., Fragoso, J. M. V., … Lavorel, S. (2003). Forecasting plant migration rates: Managing uncertainty for risk assessment. *Journal of Ecology*, *91*(3), 341–347. https://doi.org/10.1046/j.1365-2745.2003.00781.x

Hijmans, R. J., & Elith, J. (2013). *Species distribution modeling with R Introduction*. October, 71. https://doi.org/10.1016/S0550-3213(02)00216-X

Hijmans, R. J., & Elith, J. (2015). *Species distribution modeling with R Introduction*.

Horning, N. (2010). *Random Forests: An algorithm for image classification and generation of continuous fields data sets. International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences*, *2010*, 1–6. https://doi.org/10.5244/C.22.54

Ilia, I., Rozos, D., & Koumantakis, I. (2013). *Landform classification using GIS techniques. The case of Kimi municipality area, Euboea Island, Greece. Bulletin of the Geological Society of Greece, vol. XLVII 2013 Proceedings of the 13th International Congress, Chania*.

Jiménez-Valverde, A., & Lobo, J. M. (2006). The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions*, *12*, 521–524. https://doi.org/10.1111/j.1366-9516.2006.00267.x

Jiménez-Valverde, A., & Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica*, *31*(3), 361–369. https://doi.org/10.1016/j.actao.2007.02.001

Kebede, A. T., & Coppock, D. L. (2015). Livestock-mediated dispersal of *Prosopis juliflora* imperils Grasslands and the endangered Grevy's Zebra in Northeastern Ethiopia. *Rangeland Ecology and Management*, *68*(5), 402–407. https://doi.org/10.1016/j.rama.2015.07.002

Keller, R. P., Lodge, D. M., Lewis, M. A., & Shogren, J. F. (Eds.) (2009). *Bioeconomics of invasive species: Integrating ecology, economics, policy, and management*. Oxford, UK: Oxford University Press.

Kim, S. (2017). Better motivation, different thinking. https://deeplearning4j.org/neuralnet-overview.

Kimothi, M. M., & Dasari, A. (2010). Methodology to map the spread of an invasive plant (*Lantana camara* l.) in forest ecosystems using Indian remote sensing satellite data. *International Journal of Remote Sensing*, *31*(12), 3273–3289. https://doi.org/10.1080/01431160903121126

Landis, R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Le Maitre, D. C., Gush, M. B., & Dzikiti, S. (2015). Impacts of invading alien plant species on water flows at stand and catchment scales. *AoB Plants*, *7*(1), 1–21. https://doi.org/10.1093/aobpla/plv043

Lorena, A. C., Jacintho, L. F. O., Siqueira, M. F., Giovanni, R. D., Lohmann, L. G., De Carvalho, A. C. P. L. F., & Yamamoto, M. (2011). Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, *38*(5), 5268–5275. https://doi.org/10.1016/j.eswa.2010.10.031

Lu, D., Mausel, P., Brondizio, E., & Moran, E. (2002). International Journal of Assessment of atmospheric correction methods for Landsat TM data applicable to Amazon basin LBA. *International Journal of Remote Sensing*, *23*(13), 2651–2671. https://doi.org/10.1080/01431160110109642

Masocha, M., & Skidmore, A. K. (2011). Integrating conventional classifiers with a GIS expert system to increase the accuracy of invasive species mapping. *International Journal of Applied Earth Observation and Geoinformation*, *13*(3), 487–494. https://doi.org/10.1016/j.jag.2010.10.004

Menuz, D., & Kettenring, K. (2013). The importance of roads, nutrients, and climate for invasive plant establishment in riparian areas in the northwestern United States. *Biological Invasions*, *15*, 1601–1612. https://doi.org/10.1007/s10530-012-0395-6

Metz, C. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298.

Meynard, C., & Quinn, J. (2007). Predicting species distributions: A critical comparison of most common statistical models using artificial species. *Journal of Biogeography*, *34*, 1455–1469.

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose random forest to predict rare species distribution with few samples in large undersampled areas? Three Asian Crane Species Models Provide Supporting Evidence. *PeerJ*, *5*, e2849.https://doi.org/10.7717/peerj.2849

Mining, D. (2009). Springer series in statistics the elements of statistical learning. *The Mathematical Intelligencer*, *27*(2), 83–85. https://doi.org/10.1007/b94608

Mishra, N. B., Crews, K. A., & Okin, G. S. (2014). Relating spatial patterns of fractional land cover to savanna vegetation morphology using multi-scale remote sensing in the Central Kalahari. *International Journal of Remote Sensing*, *35*(6), 2082–2104.

MOA (1997). *Land resource inventory for the Afar National Regional State: Natural resource management and regulatory department*. Addis Ababa, Ethiopia: Ministry of Agriculture.

Mohamed, F. (1997). Tropical Forestry Report: Management of Prosopis juliflora for Use in Agroforestry Systems in the Sudan (Luukkanen, O. Ed.). Ph.D. Thesis, University of Helsinki, Helsinki, Finland.

MoLF (2017). *National strategy for prosopis juliflora management*. Addis Ababa, Ethiopia: Ministry of Livestock and Fisheries of Ethiopia

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Mureriwa, N., Adam, E., Sahu, A., & Tesfamichael, S. (2016). Examining the spectral separability of *Prosopis glandulosa* from co-existent species using field spectral measurement and guided regularized random forest. *Remote Sensing*, *8*(2), https://doi.org/10.3390/rs8020144

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 1–21. https://doi.org/10.3389/fnbot.2013.00021

Newcombe, R. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat Med.*, *17*, 857–872.

Ng, W., Meroni, M., Immitzer, M., Böck, S., Leonardi, U., Rembold, F., … Atzberger, C. (2016). Mapping Prosopis spp. with Landsat 8 data in arid environments: Evaluating effectiveness of different methods and

temporal imagery selection for Hargeisa, Somaliland. *International Journal of Applied Earth Observations and Geoinformation*, *53*, 76–89. https://doi.org/10.1016/j.jag.2016.07.019

Nicholls, A. O. (1989). How to make biological surveys go further with generalised linear models. *Biological Conservation*, *50*(1), 51–75. https://doi.org/10.1016/0006-3207(89)90005-0

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Olinsky, A., Kennedy, K., & Kennedy, B. (2012). Assessing gradient boosting in the reduction of misclassification error in the prediction of success for actuarial majors. *CS-BIGS*, *5*(1), 12–16.

Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, *26*(5), 1007–1011. https://doi.org/10.1080/01431160512331314083

Pal, M., & Mather, P. M. (2017). *Support vector machines for classification in remote sensing*, 1161(April). https://doi.org/10.1080/01431160512331314083

Pasiecznik, N. M., & Henry Doubleday Research Association (2001). *The* Prosopis julifora-Prosopis pallida *complex: A monograph*. Coventry, UK: HDRA.

Pearce, J., & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, *133*, 225–245.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rembold, F., Leonardi, U., Ng, W., Gadain, H., & Meroni, M. (2015). Mapping areas invaded by Prosopis juliflora in Somaliland with Landsat 8 imagery. *Proceedings of SPIE*, *9637*, 1–12. https://doi.org/10.1117/12.2193133

Riley, S., DeGloria, S., & Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, *5*, 1–4.

Rodrigues, M., & De la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling and Software*, *57*, 192–201. https://doi.org/10.1016/j.envsoft.2014.03.003

Shackleton, R. T., Le Maitre, D. C., van Wilgen, B. W. & Richardson, D. M. (2015a). The impact of invasive alien *Prosopis* species (mesquite) on native plants in different environments in South Africa. *South African Journal of Botany*, *97*, 25–31. https://doi.org/10.1016/j.sajb.2014.12.008

Shackleton, R. T., Le Maitre, D. C., van Wilgen, B. W., & Richardson, D. M. (2015b). Use of non-timber forest products from invasive alien *Prosopis* species (mesquite) and native trees in South Africa: Implications for management. *Forest Ecosystems*, *2*(1), 16. https://doi.org/10.1186/s40663-015-0040-9

Shiferaw, H., Schaffner, U., Bewket, W., Alamirew, T., Zeleke, G., Teketay, D., & Eckert, S. (2019). Modelling the current fractional cover of an invasive alien plant and drivers of its invasion in a dryland ecosystem. *Scientific Reports*, https://doi.org/10.1038/s41598-018-36587-7

Shiferaw, H., Teketay, D., Nemomissa, S., & Assefa, F. (2004). Some biological characteristics that foster the invasion of *Prosopis juliflora* (Sw.) DC. at Middle Awash Rift Valley Area, north-eastern, Ethiopia. *Journal of Arid Environments*, *58*(2), 135–154. https://doi.org/10.1016/j.jaridenv.2003.08.011

Stohlgren, T. J., Ma, P., Kumar, S., Rocca, M., Morisette, J. T., Jarnevich, C. S., & Benson, N. (2010). Ensemble habitat mapping of invasive plant species. *Risk Analysis*, *30*(2), 224–235. https://doi.org/10.1111/j.1539-6924.2009.01343.x

Tegegn, G. (2008). *Experiences on* Prosopis *management case of Afar region* (pp. 1–35). London, UK: FARM-Africa.

Tessema, Y. A. (2012). Ecological and economic dimensions of the paradoxical invasive species-Prosopis juliflora and policy challenges in Ethiopia. *Journal of Economics and Sustainable Development* (Www.liste. Org), *3*(8), 62–70.

Thessen, A. (2016). Adoption of machine learning techniques in ecology and earth science. *One Ecosystem*, *1*, e8621. https://doi.org/10.3897/oneeco.1.e8621

van Wilgen, B. W., & Wannenburgh, A. (2016). Co-facilitating invasive species control, water conservation and poverty relief: Achievements and challenges in South Africa's Working for Water programme. *Current Opinion in Environmental Sustainability*, *19*, 7–17. https://doi.org/10.1016/j.cosust.2015.08.012

Vilà, M., Espinar, J. L., Hejda, M., Hulme, P. E., Jarošík, V., Maron, J. L., … Pyšek, P. (2011). Ecological impacts of invasive alien plants: A meta-analysis of their effects on species, communities and ecosystems. *Ecology Letters*, *14*(7), 702–708. https://doi.org/10.1111/j.1461-0248.2011.01628.x

Wakie, T. T., Evangelista, P. H., Jarnevich, C. S., & Laituri, M. (2014). Mapping current and potential distribution of non-native *Prosopis juliflorain* the Afar region of Ethiopia. *PLoS ONE*, *9*(11), 3–11. https://doi.org/10.1371/journal.pone.0112854

Wakie, T., Evangelista, P., & Laituri, M. (2012). *Utilization Assessment of Prosopis juliflora in Afar Region*, Ethiopia, (July), 1–15.

Wana, D., & Beierkuhnlein, C. (2010). Plant species and growth form richness along altitudinal gradients in the southwest Ethiopian highlands. *Journal of Vegetation Science*, *21*, 617–626. https://doi.org/10.1111/j.1654-1103.2010.01177.x.

Weiss, A. (2001). *Topographic position and landforms analysis. Poster presentation, ESRI User Conference*. San Diego, CA.

Wilson, J. R. U., Dormontt, E. E., Prentis, P. J., Lowe, A. J., & Richardson, D. M. (2009). Something in the way you move: Dispersal pathways affect invasion success. *Trends in Ecology and Evolution*, *24*(3), 136–144. https://doi.org/10.1016/j.tree.2008.10.007

Yirgalem, A. (2001). *Challenges, opportunities and prospects of common property resources management in the Afar Regional Areas*. Addis Abeba, Ethiopia: FARM-Africa, Mimeo.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., & Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region. Saudi Arabia. *Landslides*, *13*(5), 839–856. https://doi.org/10.1007/s10346-015-0614-1

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 22–40. https://doi.org/10.1109/MGRS.2016.2540798

Zhou, Y., Chen, J., Cao, X., & Chen, X. (2015). Estimation of fractional vegetation cover in semiarid areas by integrating endmember reflectance purification into nonlinear spectral mixture. *Analysis*, *12*(6), 1175–1179.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Shiferaw H, Bewket W, Eckert S. Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem. *Ecol Evol*. 2019;00:1–13. https://doi.org/10.1002/ece3.4919